NASHVILLE SOFTWARE SCHOOL

File Types

Data engineering

Files types we will cover

- .csv
- .parquet
- .json
- .xml

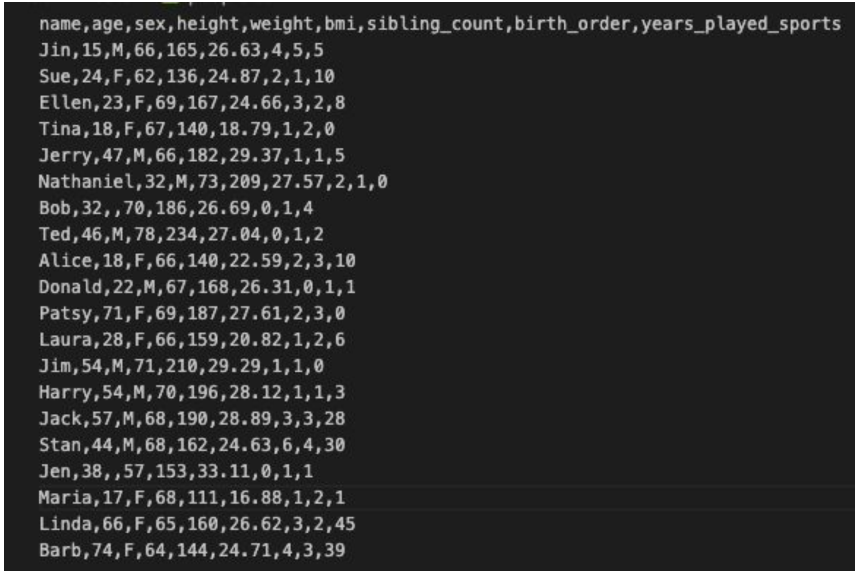
.CSV

Pros:

- Easy to understand and edit
- Compatible with many different applications
- Lightweight and compact

Cons:

- Prone to user error
- Formatting issues if opened in spreadsheet software (Excel drops leading zeros)
- High memory usage at scale



```
name,age,sex,height,weight,bmi,sibling_count,birth_order,years_played_sports
Jin,15,M,66,165,26.63,4,5,5
Sue,24,F,62,136,24.87,2,1,10
Ellen,23,F,69,167,24.66,3,2,8
Tina,18,F,67,140,18.79,1,2,0
Jerry,47,M,66,182,29.37,1,1,5
Nathaniel,32,M,73,209,27.57,2,1,0
Bob,32,,70,186,26.69,0,1,4
Ted,46,M,78,234,27.04,0,1,2
Alice,18,F,66,140,22.59,2,3,10
Donald,22,M,67,168,26.31,0,1,1
Patsy,71,F,69,187,27.61,2,3,0
Laura,28,F,66,159,20.82,1,2,6
Jim,54,M,71,210,29.29,1,1,0
Harry,54,M,70,196,28.12,1,1,3
Jack,57,M,68,190,28.89,3,3,28
Stan,44,M,68,162,24.63,6,4,30
Jen,38,,57,153,33.11,0,1,1
Maria,17,F,68,111,16.88,1,2,1
Linda,66,F,65,160,26.62,3,2,45
Barb,74,F,64,144,24.71,4,3,39
```

.parquet

Pros

- Efficient compression and fast query performance
- Columnar file format that is highly optimized for big data reads

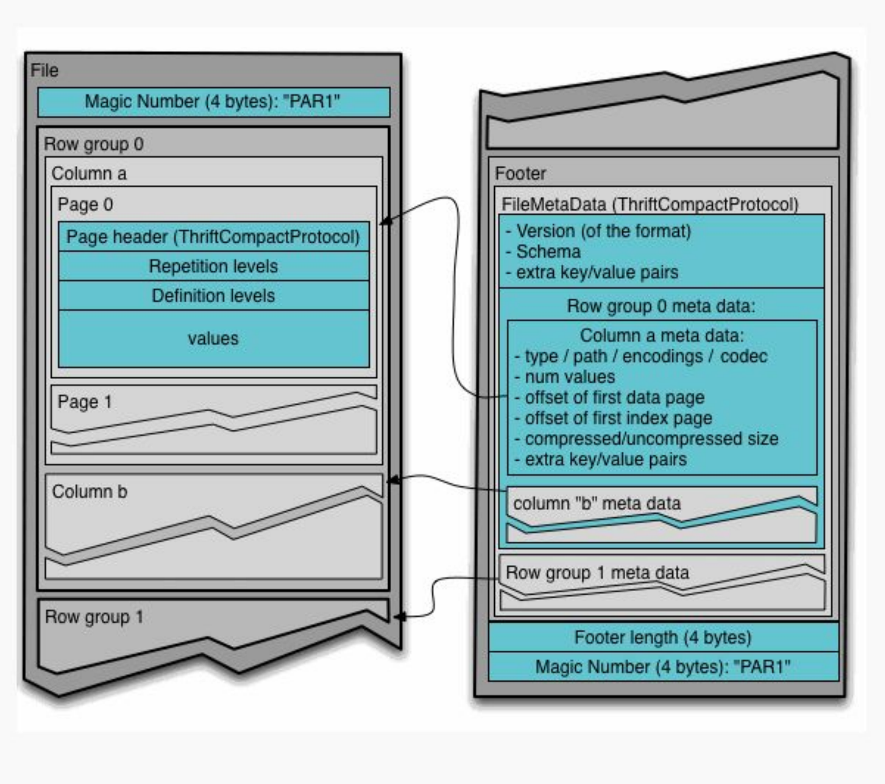
Cons:

- Slower write performance.

.parquet cont

Structure of a Parquet File

- Header
  - Contains the initial magic number “PAR1” (What is a magic number?)
- Data Blocks
  - Where the data is stored.
  - Columnar format split into row groups with data stored in a column chunk.
- Footer
  - File metadata (describes the schema)
  - Footer length 4 bytes
  - Magic Number: “PAR1” indicating the end of the file.



File
Magic Number (4 bytes): "PAR1"
Row group 0
Column a
Page 0
Page header (ThriftCompactProtocol)
Repetition levels
Definition levels
values
Page 1
Column b
Row group 1
Footer
FileMetaData (ThriftCompactProtocol)
- Version (of the format)
- Schema
- extra key/value pairs
Row group 0 meta data:
Column a meta data:
- type / path / encodings / codec
- num values
- offset of first data page
- offset of first index page
- compressed/uncompressed size
- extra key/value pairs
column "b" meta data
Row group 1 meta data
Footer length (4 bytes)
Magic Number (4 bytes): "PAR1"

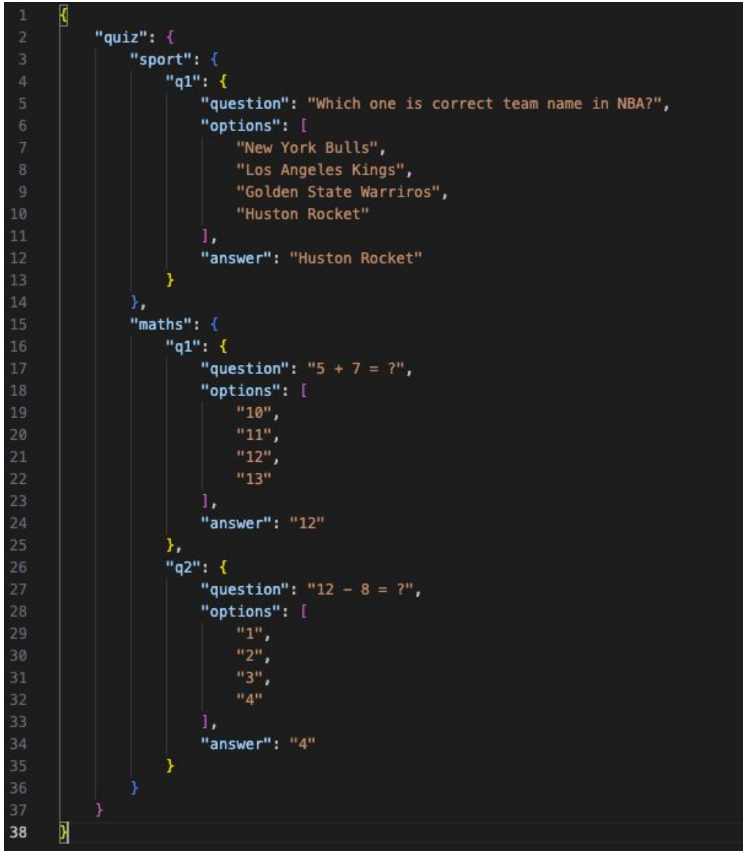
.json

Pros

- Human readable
- Lightweight and efficient
- White Support

Cons:

- No built in schema validation
- complexity with nested data



```json
{
    "quiz": {
        "sport": {
            "q1": {
                "question": "Which one is correct team name in NBA?",
                "options": [
                    "New York Bulls",
                    "Los Angeles Kings",
                    "Golden State Warriros",
                    "Huston Rocket"
                ],
                "answer": "Huston Rocket"
            }
        },
        "maths": {
            "q1": {
                "question": "5 + 7 = ?",
                "options": [
                    "10",
                    "11",
                    "12",
                    "13"
                ],
                "answer": "12"
            },
            "q2": {
                "question": "12 - 8 = ?",
                "options": [
                    "1",
                    "2",
                    "3",
                    "4"
                ],
                "answer": "4"
            }
        }
    }
}
```

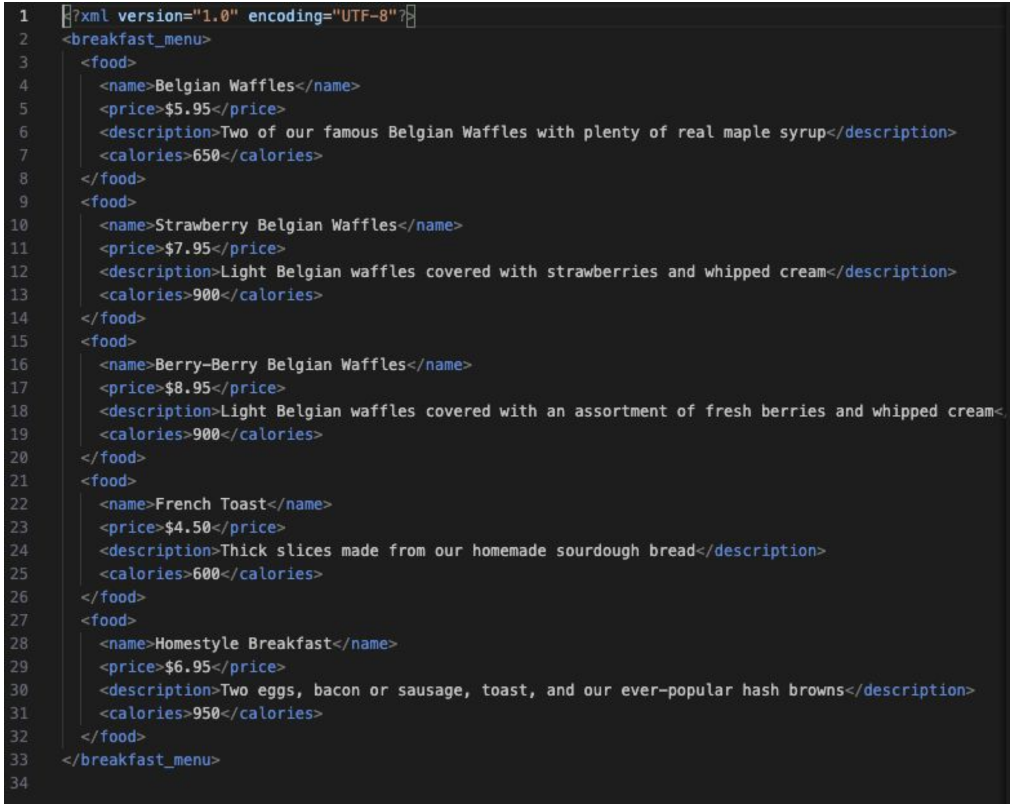
.xml

Pros

- Platform independent
- Allows validation

Cons

- XML is verbose compared to other formats
- Added transportation cost due to redundancy
- Less readable
- Doesn't support arrays
- Files can be large



```xml
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
  <food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    <description>Two of our famous Belgian Waffles with plenty of real maple syrup</description>
    <calories>650</calories>
  </food>
  <food>
    <name>Strawberry Belgian Waffles</name>
    <price>$7.95</price>
    <description>Light Belgian waffles covered with strawberries and whipped cream</description>
    <calories>900</calories>
  </food>
  <food>
    <name>Berry-Berry Belgian Waffles</name>
    <price>$8.95</price>
    <description>Light Belgian waffles covered with an assortment of fresh berries and whipped cream<
    <calories>900</calories>
  </food>
  <food>
    <name>French Toast</name>
    <price>$4.50</price>
    <description>Thick slices made from our homemade sourdough bread</description>
    <calories>600</calories>
  </food>
  <food>
    <name>Homestyle Breakfast</name>
    <price>$6.95</price>
    <description>Two eggs, bacon or sausage, toast, and our ever-popular hash browns</description>
    <calories>950</calories>
  </food>
</breakfast_menu>
```

Summary

We talked about:

- CSV
- Parquet
- JSON
- XML

Which one you will use will depend on the nature of the data, what the current system accepts, and what format fits the architecture you are developing.